

## PATENT ABSTRACTS OF JAPAN

(11)Publication number : 11-149350

(43)Date of publication of application : 02.06.1999

(51)Int.Cl.

G06F 3/06

G06F 3/06

G06F 12/08

G06F 12/08

(21)Application number : 10-174327

(22)Date of filing : 22.06.1998

(71)Applicant : HITACHI LTD

(72)Inventor : YAMAMOTO AKIRA

SATO TAKAO

HONMA SHIGEO

AZUMI YOSHIHIRO

KUWABARA YOSHINAGA

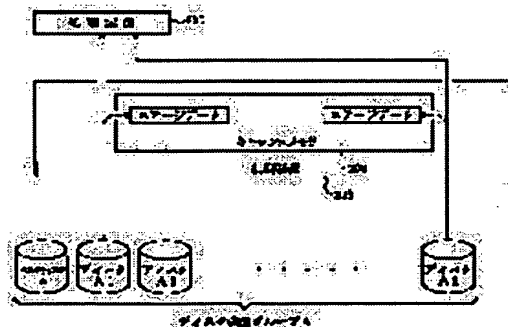
KITAJIMA HIROYUKI

## (54) DISK STORAGE SYSTEM

(57)Abstract:

**PROBLEM TO BE SOLVED:** To decentralize the load of input/output processing which should be performed between disk devices in a disk device group and to improve the parallelism of the execution of the input/output processing by selecting some arbitrary disk device in a free state when a read request which needs to access a disk device is received.

**SOLUTION:** A controller 203 processes the read request which is received from a processor 210 and needs to access the disk group A with a disk A1. For example, the controller 203 receives the read request which needs to access the disk device group A in this state from the processor 210. The controller 203 selects an arbitrary disk device in a free state in the disk device group A, i.e., a disk Ai and starts processing the received read request. Even when a write request which needs to access the disk device group A is received, the disk device in the free state starts processing it.



## LEGAL STATUS

[Date of request for examination]

[Date of sending the examiner's decision of rejection]

[Kind of final disposal of application other than the examiner's decision of rejection or application converted registration]

[Date of final disposal for application]

[Patent number]

[Date of registration]

[Number of appeal against examiner's decision  
of rejection]

[Date of requesting appeal against examiner's  
decision of rejection]

[Date of extinction of right]

Copyright (C); 1998,2000 Japanese Patent Office

(19) 日本国特許庁 (J P)

(12) 公開特許公報 (A)

(11) 特許出願公開番号

特開平11-149350

(43) 公開日 平成11年(1999) 6月2日

(51) Int.Cl.<sup>6</sup>

G 0 6 F 3/06

12/08

識別記号

3 0 2

3 0 4

3 2 0

F I

G 0 6 F 3/06

12/08

3 0 2 Z

3 0 4 E

B

J

3 2 0

審査請求 有 請求項の数 3 O L (全 26 頁)

(21) 出願番号

特願平10-174327

(62) 分割の表示

特願平2-42452の分割

(22) 出願日

平成2年(1990) 2月26日

(71) 出願人 000005108

株式会社日立製作所

東京都千代田区神田駿河台四丁目6番地

(72) 発明者 山本 彰

神奈川県川崎市麻生区王禅寺1099番地 株

式会社日立製作所システム開発研究所内

(72) 発明者 佐藤 孝夫

神奈川県川崎市麻生区王禅寺1099番地 株

式会社日立製作所システム開発研究所内

(72) 発明者 本間 繁雄

神奈川県小田原市国府津2880番地 株式会

社日立製作所小田原工場内

(74) 代理人 弁理士 小川 勝男

最終頁に続く

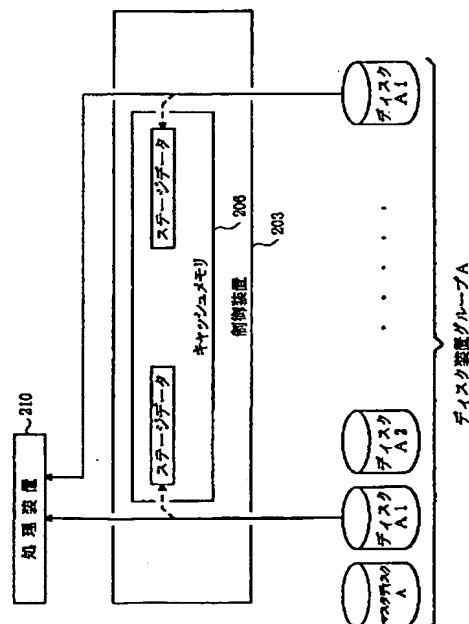
(54) 【発明の名称】 デイスクストレージシステム

(57) 【要約】

【課題】 記憶装置グループに対する複数のリード要求を効率良く実行する。

【解決手段】 制御装置は、第1のリード要求と第2のリード要求を、記憶装置グループ内の異なる記憶装置と、異なるバスと、異なる外部コネクションポイントを用いて実行する。

図5



## 【特許請求の範囲】

【請求項1】複数の外部接続ポイントを有する制御装置と、複数の記憶装置から構成されライトデータが各記憶装置に書き込まれる記憶装置グループと、各々が前記記憶装置グループと前記制御装置とに接続される複数のバスとから構成されるディスクストレージシステムであって、

前記制御装置は、

前記記憶装置グループに対する第1と第2のリード要求を受け取り、

前記第1のリード要求で要求されたデータを前記記憶装置グループ内の前記複数のうちのいずれか1つのバスにより接続されたいずれか1つの記憶装置から読み出す第1のリードオペレーションと、前記第2のリード要求で要求されたデータを前記記憶装置グループ内の前記複数のうちの他の1つのバスにより接続された他の1つの記憶装置から読み出す第2のリードオペレーションを実行し、前記第1のリードオペレーションで読み出されたデータを前記複数のうちのいずれか1つの外部接続ポイントに転送する第1の転送オペレーションと、前記第2のリードオペレーションで読み出されたデータを前記複数のうちの他の1つの外部接続ポイントに転送する第2の転送オペレーションを実行することを特徴とするディスクストレージシステム。

【請求項2】前記第1と第2のリード要求は、処理装置から前記記憶装置グループに発行されたものであって、前記制御装置は、前記第1のリードオペレーションで読み出したデータを前記いずれか1つの外部接続ポイントを介して前記処理装置に転送し、前記第2のリードオペレーションで読み出したデータを前記他の1つの外部接続ポイントを介して前記処理装置に転送することを特徴とする請求項1に記載のディスクストレージシステム。

【請求項3】前記制御装置は、第1の処理装置に接続可能な第1の外部接続ポイントと、第2の処理装置に接続可能な第2の外部接続ポイントとを備え、

前記制御装置は、前記第1の処理装置から発行された前記第1のリード要求と、前記第2の処理装置から発行された前記第2のリード要求とを受け取り、

前記処理装置は、前記第1のリードオペレーションで読み出されたデータを前記第1の処理装置に前記第1の外部接続ポイントを介して転送し、前記第2のリードオペレーションで読み出されたデータを前記第2の処理装置に前記第2の外部接続ポイントを介して転送することを特徴とする請求項1に記載のディスクストレージシステム。

## 【発明の詳細な説明】

## 【0001】

【発明の属する技術分野】本発明は、記憶装置の間で、

実行すべき負荷を分散するためのディスクストレージシステムに関する。

## 【0002】

【従来の技術】特開昭60-114947では、ディスクキャッシュ（以下、単にキャッシュと呼ぶ場合がある。）を有する制御装置におけるディスク装置の2重書き機能に関する技術が開示されている。ここでは、同一のデータを書き込む2つのディスク装置を、2重書きディスクと呼ぶ。

【0003】特開昭60-114947には、ディスクキャッシュを利用してライトデータ2重（2つの）のディスク装置を制御する方法が、以下のように説明されている。制御装置は、CPUからの入出力要求を、2つのディスク装置のうちの1つのディスク装置に対して処理する。リード要求（入力要求）を受け付けた場合には、制御装置は受け付けた要求をそのまま実行する。CPUからライト要求（出力要求）を受け付けた場合には、以下の処理を実行する。すなわち、2重書きディスクのうちの一方のディスク装置に、データを書き込むと共に、ディスクキャッシュにもこのデータを書き込んでおく。そして、制御装置は後で、空いた時間を利用して、ディスクキャッシュに書き込んでおいたデータを他方のディスク装置に書き込む。この書き込み処理は、ライトアフタ処理と呼ばれる。以上により、2重書きディスク装置に同一のデータが書き込まれる。

【0004】特公昭61-28128では、2重化ファイルの制御方法として2重書きディスク装置の負荷分散に関する技術が開示されている。特公昭61-28128は、入出力要求を処理する際、空いているディスク装置を選択することにより処理の高速化をねらったものである。ただし、特公昭61-28128は、特開昭60-114947のように、ディスクキャッシュを利用したライトアフタ制御によりデータを2重書きする方法を用いていない。

【0005】西垣他：順次アクセス入力処理におけるディスクキャッシュの効果解析、情報処理学会論文誌、V o l . 25, N o . 2, p p . 313-320 (1984) の論文においては、以下の技術が開示されている。ディスクキャッシュを有する制御装置における、シーケンシャル処理に対する先読み処理、すなわち、CPUからは要求されていないデータを、キャッシュにステージする技術である。このステージ処理は、CPUからの入出力要求に対応するとは、独立に制御装置が実行する。

## 【0006】

【発明が解決しようとする課題】特開昭60-114947は、2重書きディスクにおいては制御装置が処理対象とすることができるディスク装置は複数存在するという利点に注目していない。すなわち、CPUから受け付けた入出力要求に対して選択すべきディスク装置を特定の1台に限定している。

【0007】一方、特公昭61-28128で開示されているように、入出力要求に対応して、空いているディスク装置を選択するという方法は、性能的には優れている。しかし、この方法を、キャッシュを利用したライトアフタ制御により2重書き機能を実現するケースに適用すると信頼性の点で劣化が生ずる。これは、すべてのディスク装置に対応するCPUから受け付けた末書き込みのライトデータがある可能性が高いためである。したがって、キャッシュの電源のダウンと、どれか1台のディスク装置の障害が重なると、CPUから受け付けたライトデータが消失することになる。

【0008】さらに、キャッシュを有する制御装置の場合、CPUから受け取る入出力要求とは独立に、制御装置がキャッシュとディスク装置との間で、入出力処理を実行する。(これは、西垣他：順次アクセス入力処理におけるディスクキャッシュの効果解析、情報処理学会論文誌、Vol. 25, No. 2, pp. 313-320 (1984)に開示されている。)したがって、制御装置がCPUからの入出力要求とは、独立に実行する入出力処理に対しても、その処理を割り当てるべき対象として選択できるディスク装置が複数存在するという点に注目すべきである。

【0009】本発明は、キャッシュを利用したライトアフタ処理により、1台以上のディスク装置からなるディスク装置グループに対して同一のデータを書き込む制御方法および制御装置に関する。

【0010】本発明の目的は、ディスク装置グループ内のディスク装置との間で実行すべき入出力処理の負荷を分散し、入出力処理の実行の並列度を向上させる制御装置および制御方法を提供することにある。

【0011】

【課題を解決するための手段】課題を解決するための方法および装置の説明を分かり易くするために、制御装置がディスク装置との間で実行する入出力処理を分類する。

【0012】制御装置がディスク装置との間で実行する入出力処理は、以下の4種類に分類できる。

【0013】(1) CPUから受け付けたライト要求に対する処理であって、ディスク装置にアクセスする必要がある処理。

【0014】(2) CPUから受け付けたリード要求に対する処理であって、ディスク装置にアクセスする必要がある処理。

【0015】(3) CPUからの入出力要求(リード要求/ライト要求)とは独立した、ディスク装置からキャッシュへデータを転送するステージ処理。

【0016】(4) ディスク装置との間で実行するライトアフタ処理。

【0017】以上の入出力処理のうち、ライトアフタ処理は負荷分散の対象とはならない。以下、その理由を述

べる。ライトアフタ処理は、CPUから受け付けたライト要求の実行の際に、データを書き込んだディスク装置以外のディスク装置グループ内のすべてのディスク装置に対して実行する処理である。このため、ライトアフタ処理を実行すべきディスク装置を選択する自由度はないことになる。したがって、上記の(1)から(4)の処理の内、(1)から(3)の処理が負荷分散の対象となる。

【0018】本明細書では、負荷分散を実行する方法として2種類の方法を示す。便宜的に、これらの方法を負荷分散方法1、負荷分散方法2と呼ぶ。以下、それぞれの方法について述べる。

【0019】負荷分散方法1…制御装置は、上記(2)または(3)の入出力処理を実行すべきディスク装置を選択する際、ディスク装置グループ内の任意の空いた状態にあるディスク装置を選択する。上記(1)のディスク装置にアクセスする必要があるCPUから受け付けたライト要求に対応してディスク装置を選択する際は、ディスク装置グループ内の特定のディスク装置を選択する。

【0020】負荷分散方法2…制御装置は、前述の処理分類(1)、すなわち、ディスク装置にアクセスする必要があるCPUから受け付けたライト要求に対応して、ディスク装置を選択する際は、ディスク装置グループ内の特定のディスク装置を選択する。処理分類(2)または(3)に示した入出力処理の実行すべきディスク装置を選択する際は、上記の特定のディスク装置以外のディスク装置を優先的に選択する。

【0021】負荷分散方法1および負荷分散方法2のそれぞれに対応した作用について述べる。

【0022】(1) 負荷分散方法1の場合  
制御装置は、あるディスク装置グループに対して、ディスク装置にアクセスする必要があるリード要求をCPUから受け取ると、以下の処理を実行する。そのリード要求に対して、ディスク装置グループ内の空いた状態にある任意の1台のディスク装置を選択する。空いた状態のディスク装置が1台もない場合、制御装置はそのリード要求を待たせる。また、ディスク装置にアクセスする必要があるライト要求を受け取った場合には、そのライト要求に対して、ディスク装置グループ内の特定の1台のディスク装置を選択する。ただし、この特定のディスク装置が空いた状態になれば、制御装置はそのライト要求を待たせる。

【0023】また、制御装置が、CPUからの入出力要求とは独立したステージ処理を実行しようとした場合にも、以下のような処理を行う。すなわち、このステージ処理に対応し、ディスク装置グループ内の空いた状態にある任意の1台のディスク装置を選択する。空いた状態のディスク装置が1台もない場合、制御装置は、そのステージ処理を実行しない。

【0024】負荷分散方法1の信頼性および性能面の特徴を以下に示す。

【0025】負荷分散方法1は、特開昭60-114947に開示されている方法に比べて、分散の効果はやや劣るものの、特公昭61-28128に開示されている方法に較べると、優れた性能を得ることができる。

【0026】以下、この理由を述べる。負荷分散方法1は、ライト要求に対するディスク装置選択の自由度に制限を設けていることになる。したがって、特開昭60-114947に開示されている方法に比べて、分散の効果は劣ることになる。ただし、リード要求に対しては空いた状態にあるディスク装置を選択する。通常、ディスク装置に対する入出力要求の割合は、リード要求の方がライト要求に比べて、かなり多い。(3対1から4対1程度)したがって、負荷分散方法1は、特開昭60-114947に開示されている方法に較べ、それほど性能劣化は生じないことになる。一方、すべての入出力要求を1台のディスク装置に集中させる特公昭61-28128の方法に比べると、負荷分散方法1は優れた性能を得ることができる。

【0027】負荷分散方法1の信頼性は、特開昭60-114947に開示されている方法に比べて高く、特公昭61-28128の方法に較べると、ほぼ同等である。

【0028】以下、その理由を述べる。負荷分散方法1、または、特公昭61-28128では、ライト要求を集中させるディスク装置に対して、ライトアプタ処理すべきデータがない。したがって、キャッシュの電源がダウンしても、ライト要求を集中させるディスク装置に障害が発生しなければ、CPUから受け付けたライトデータが消失しない。

【0029】以上より、負荷分散方法1により、ディスク装置グループの高性能化/高信頼化をバランスよく実現する負荷分散が可能となる。

【0030】(2)負荷分散方法2の場合  
制御装置は、ディスク装置グループのディスク装置にアクセスする必要のあるライト要求をCPUから受け取った場合には、そのライト要求に対応して、ディスク装置グループ内の特定の1台のディスク装置を選択する。ただし、この特定のディスク装置が空いた状態になれば、制御装置は、そのライト要求を待たせる。また、あるディスク装置グループのディスク装置にアクセスする必要のあるリード要求をCPUから受け取ると、以下の処理を実行する。まず、上記の特定のディスク装置以外のディスク装置の中で、空いた状態にある任意の1台のディスク装置を選択する。空いたディスク装置が、1台もないとき、上記の特定のディスク装置が空いているか否かを調べる。空いている場合、制御装置は、この特定のディスク装置を、受け取ったリード要求のために選択する。空いていない場合、制御装置は、そのリード要求

を待たせる。

【0031】また、制御装置が、CPUからの入出力要求とは独立したステージ処理を実行しようとした場合にも、以下のような処理を行う。まず、上記の特定のディスク装置以外のディスク装置の空いた状態にある任意の1台のディスク装置を選択する。空いたディスク装置が、1台もない場合、上記の特定のディスク装置が空いているか否かを調べる。空いている場合、制御装置は、この特定のディスク装置を、実行しようとしたステージ処理のために選択する。空いていない場合、制御装置は、そのステージ処理を実行しない。

【0032】以上述べた分散方法をとる理由は、以下のとおりである。例えば、CPUからのライト要求に対応する処理を集中させる特定のディスク装置に、リード要求に対応する処理割り当てるとする。この場合、リード要求に対応する処理が完了しないうちに、ライト要求を受け取ると、ライト要求に対応する処理に入れないことになる。したがって、特定のディスク装置を選択する必要のないCPUからのライト要求に対応する処理以外の処理は、この特定のディスク装置以外のディスク装置を優先的に割り当てをした方が、分散の効果を高くすることができる。

【0033】

【発明の実施の形態】以下、本発明について、2種類の実施例を説明するが、まず、両実施例に共通する内容について述べる。

【0034】図2は、本発明の適用対象となる計算機システムの構成である。計算機システムは、CPU200、主記憶201、チャンネル202とからなる処理装置210と、制御装置203と、1台以上のn台のディスク装置204とより構成される。なお、制御装置203に複数の処理装置210が接続されている場合にも、本発明を適用できることは後述の内容から明らかになる。

【0035】n台のディスク装置204は、それぞれ1台以上のディスク装置204からなるm個のディスク装置グループ211にグループ化されている。各ディスク装置グループ211に属するディスク装置204の台数は各々異なってもよい。それぞれのディスク装置204は、ある1つのディスク装置グループ211に属する。それぞれのディスク装置204が、どのディスク装置グループ211に属するかを決定する方法は、本発明には直接関係ないため、説明を省略する。

【0036】制御装置203は、1つ以上のディレクタ205、キャッシュ(メモリ)206、制御情報用メモリ207およびディレトリ208より構成される。各ディレクタ205は、チャンネル202とディスク装置204との間、チャンネル202とキャッシュ206との間、ならびに、キャッシュ206とディスク装置204との間でデータを転送する。キャッシュ206には、ディスク装置204に格納されているデータの中でアクセ

ス頻度の高いデータをステージしておく。ディレクトリ208は、キャッシュ206の管理情報を格納する。ステージ処理は、ディレクトリ205によって実行される。具体的なステージデータの例は、CPU20からのアクセス対象となったデータ、および、このデータとディスク装置204の格納位置に近いデータなどである。

【0037】本発明の対象となる制御装置203は、あるディスク装置グループ211に属するディスク装置204に同一のデータを書き込む機能、いわゆる、多重書き機能を有する。したがって、処理装置210は、それぞれのディスク装置グループ211に対して入出力要求を発行すると考えてよい。

【0038】本発明においては、制御装置203から見ると、処理装置204から受け付ける入出力要求は以下のように分類できる。

【0039】(1) キャッシュ206と処理装置210との間のデータ転送の要求であり、ディスク装置グループ211には、アクセスしない入出力処理パターンである。これは、例えば、処理装置210から受け付けたリード要求に対応するデータがキャッシュ206内にステージされている場合に実行される処理パターンである。

【0040】(2) ディスク装置グループ211内のディスク装置204にアクセスが必要となる入出力処理パターンである。

【0041】(3) さらに、制御装置203が、キャッシュ206を有する場合、処理装置210から受け付けた入出力要求とは独立に、制御装置203が以下の入出力処理を実行する。

【0042】キャッシュ206とディスク装置グループ211内のディスク装置204との間の入出力処理パターンであり、処理装置210が関与しないデータ転送パターンである。

【0043】本発明は、同一のディスク装置グループ211に属するディスク装置204の間の負荷分散方法に関する。したがって、(1)に示したディスク装置グループ211にアクセスする必要のない処理パターンは、本発明には、直接関係しないことになる。制御装置203が実行する入出力処理パターンのうち(2)および(3)に示した入出力処理パターンが本発明の対象となることになる。なお、(2)あるいは(3)に示した入出力処理を割り当てられていない(処理を実行中でない)ディスク装置204を、空いた状態にあるディスク装置と呼ぶ。

【0044】以下、2種類それぞれの実施例の概要を説明する。まず、第1の実施例の概要を説明する。

【0045】図1は、第1の実施例における、制御装置203の動作を説明する図である。

【0046】図1の構成においては、それぞれのディスク装置グループA、BおよびC内に、1台のマスタディスクA、B、およびCが存在する。マスタディスクと他

のディスク装置との相違については、後述する。マスタディスクとは、各ディスク装置グループ内で予め定めた特定のディスク装置である。

【0047】図1においては、ディスク装置グループAおよびディスク装置グループBおよびディスク装置グループCが制御装置203に接続されている。

【0048】ディスク装置グループAは、マスタディスクA、ディスクA1、…、ディスクAiによって構成される。同様に、ディスク装置グループBは、マスタディスクB、…、ディスクBjによって、ディスク装置グループCは、マスタディスクC、…、ディスクCkによってそれぞれ構成される。

【0049】制御装置203が、処理装置210から受け取り、ディスク装置グループ211にアクセスする必要のある入出力要求を、ライト要求とリード要求に分けて説明する。図1において、ライト要求に伴うデータの流れを符号110、リード要求に伴うデータの流れを符号113で示してある。以下、それぞれの要求に対するデータ転送パターンについて述べる。

【0050】制御装置203は、ディスク装置グループAにアクセスする必要のあるライト要求110を受け取っている。制御装置203は、ディスク装置グループAの中で、マスタディスクAを選択する。すなわち、マスタディスクAとは、ディスク装置グループAにアクセスする必要のあるライト要求を集中させるディスク装置ということになる。

【0051】制御装置203は、処理装置210から受け付けたライト要求に伴うデータをマスタディスクAに書き込むと共に、キャッシュ206にも書き込む。制御装置203は、同一のディスク装置グループに属するディスク装置、例えば、マスタディスクA、ディスク装置A1、…、ディスク装置Aiには、同一のデータの書き込み処理を実行する。キャッシュ206内に書き込んだライトデータ111は上記の書き込み処理を実行するために使用する。具体的には、後で、説明する。

【0052】ディスク装置グループAにアクセスする必要のあるライト要求を受け付けた場合、マスタディスクAを選択する理由は、以下のとおりである。ディスク装置グループAにアクセスする必要のあるライト要求をマスタディスクAに必ず割り当てるとすると、マスタディスクAには、処理装置210から受け取ったライトデータ112のすべてを書き込むことになる。このため、例えば、マスタディスクA以外のディスクAiが故障して、キャッシュ206が電源ダウンした場合にも、マスタディスクAに完全なデータが保持される。

【0053】しかし、処理装置210から受け付けたライト要求の割当てに、ディスク装置の選択の自由度を制限しているため性能は落ちることになる。具体的には、制御装置203が、ディスク装置グループAにアクセスする必要のあるライト要求110を受け付けた時、マス

タディスクA以外のディスクA1などが空いていても、マスタディスクAが空いていない場合、直ちにその処理に入ることができないことになる。

【0054】図1では、制御装置203は、ディスク装置処理グループCにアクセスする必要のあるリード要求を受け取った場合も示している。この時、制御装置203は、ディスク装置グループCの中で、空いた状態にある任意のディスク装置の中から1つのディスク装置(図1では、ディスクC1)を選択する。

【0055】制御装置203は、ディスクC1から処理装置210に対して、要求されたデータを送る。(この経路を符号113で示す。)この時、処理装置210が要求したデータ、(もしくは、処理装置210が要求したデータとこのデータのディスクC1上の近傍のデータ)をステージデータ114として、キャッシュ206にステージしてもよい。この様子を破線でしめす。

【0056】制御装置203が、処理装置210から受け付けた入出力要求とは独立に実行するディスク装置グループとキャッシュ206との間の入出力処理は、図1に示した以下の処理がある。ライトデータを、ディスク装置に書き込むライトアフト処理と、処理装置210からの入出力要求とは独立なステージ処理である。処理装置210とは独立なステージ処理の例は、シーケンシャル処理に対する先読み処理を制御装置203が実行する場合に相当する。以下、それぞれの場合について説明する。

【0057】ライトアフト処理は、キャッシュ206内のライトデータ111を、未書き込みであるディスク装置Aiに、書き込む処理である。ディスク装置グループAの中で、処理装置210から受け取ったライトデータ111を直接書き込んだマスタディスクAには、ライトアフト処理を実行する必要はない。したがって、マスタディスクAを除く、ディスクA1, ..., ディスクAiには、順次ライトアフト処理が実行される。実行順序は不同でもよい。

【0058】制御装置203は、あるディスク装置グループBに対して処理装置210とは独立な、ステージ処理を実行する場合、ディスク装置グループBの中で空いた状態にある任意の1台のディスク装置を選択する。図1では、制御装置203は、ディスクBjから、ステージデータ116をキャッシュ206にステージしている。

【0059】以上が、本実施例における制御装置203の動作である。その特徴点は、以下のとおりである。

【0060】処理装置210から受け付けたディスク装置グループにアクセスする必要があるライト要求に対しては、マスタディスクを選択して信頼性を確保する。一方、ディスク装置グループからデータを読み出す場合、空いた状態にあるディスク装置を選択する。以上により、高信頼化と高性能化とをバランスよく実現する。

【0061】以下、図3から図5を用いて、並列動作の例を示す。

【0062】図3は、制御装置203が、以下のパターンの入出力処理の並列実行の形態を表している。

【0063】第1のパターンの入出力処理は、処理装置210から受け付けたディスク装置グループAにアクセスする必要のある入出力処理である。

【0064】第2のパターンの入出力処理は、処理装置210とは独立に制御装置203がキャッシュ206との間で実行する入出力処理である。

【0065】例えば、図3に示すように、制御装置203が、ディスクA1との間で処理装置210からの入出力処理とは独立なステージ処理、ディスクA2との間でライトアフト処理を実行しているとする。この時、さらに、制御装置203が、処理装置210からディスク装置グループAへのアクセスが必要な入出力要求を受け取ったとする。図3は、ディスク装置グループAへのアクセスが必要なリード要求を受け取った例である。この場合、制御装置203は、ディスク装置グループAの中から空いた状態のディスク装置Aiを選択することにより、受け取った入出力要求の実行に入ることができる。ただし、処理装置210から受け取った入出力要求が、ディスク装置グループAへのアクセスが必要なライト要求の場合には、マスタディスクAが空いていないと、同様の処理が実行できない。

【0066】また、図3では、それぞれライトアフト処理、処理装置210とは独立なステージ処理を1多重ずつ実行している。ただし、空いた状態にあるディスク装置があれば、制御装置203は、それぞれ2多重以上のライトアフト処理、処理装置210とは独立なステージ処理を実行可能である。

【0067】図4および図5は、処理装置210から受け付けた入出力処理の並列動作に関する内容である。

【0068】図4は、制御装置203に処理装置210が複数、具体的には、処理装置210と処理装置210aが接続されている場合である。

【0069】制御装置203が、処理装置210から受け取ったディスク装置グループAにアクセスする必要のあるリード要求の処理を、ディスクA1との間で処理を実行しているとする。この時、制御装置203が、処理装置210aからディスク装置グループAにアクセスする必要のあるリード要求を、受け取ったとする。制御装置203は、ディスク装置グループAの中で空いた状態にある任意のディスク装置、すなわち、ディスクAiを選択し、受け取ったリード要求の処理に入る。もちろん、処理装置210aからディスク装置グループAにアクセスする必要のあるライト要求を受け取った場合にも空いた状態にあるディスク装置があれば、直ちにその処理に入ることができる。ただし、マスタディスクAの競合により、ディスク装置グループAにアクセスする必要



のあるライト要求どうしの並列動作は実行不可能である。

【0070】さらに、3台以上の処理装置が制御装置203に接続されている場合には、空いた状態にあるディスク装置さえあれば、3多重以上のディスク装置グループAにアクセスする必要のある入出力要求を並列に実行可能である。

【0071】図5は、制御装置203に処理装置210が1台接続されている場合である。処理装置210は、ディスク装置グループAへの入出力要求に対する処理が完了する前に、ディスク装置グループAに新たな入出力要求を発行できる機能をもっているものとする。

【0072】図5において、制御装置203が、処理装置210から受け取ったディスク装置グループAにアクセスする必要のあるリード要求の処理を、ディスクA1との間で実行しているとする。この状態で、制御装置203が、処理装置210から、さらに、ディスク装置グループAにアクセスする必要のあるリード要求を受け取ったとする。制御装置203は、ディスク装置グループAの中で空いた状態にある任意のディスク装置、すなわち、ディスクAiを選択し、受け取ったリード要求の処理に入る。もちろん、処理装置210からディスク装置グループAにアクセスする必要のあるライト要求を受け取った場合にも空いた状態にあるディスク装置があれば、直ちにその処理に入ることができる。ただし、マスタディスクAの競合の関係から、ディスク装置グループAにアクセスする必要のあるライト要求どうしの並列動作は実行できない。

【0073】さらに、処理装置210が同一のディスク装置グループAに対し、3多重以上の入出力要求を発行する場合にも、空いた状態にあるディスク装置さえあれば、制御装置203は、発行された入出力要求を並列に実行可能である。

【0074】以上が、第1の実施例の概要である。次に、第2の実施例について概要について述べる。

【0075】第2の実施例の概要を、図6に示す。

【0076】第2の実施例は、制御装置203の動作が、第1の実施例と以下の点で異なる。ディスク装置グループにアクセスする必要のあるリード要求に関しては、マスタディスク以外のディスク装置を優先的に選択する点である。同様に、処理装置210とは独立なステージ処理に関しても、マスタディスク以外のディスク装置を優先的に選択する。以上述べた方法をとる理由は、マスタディスク以外のディスク装置を選択しておけば、さらなる高速化が可能となるためである。というのは、ディスク装置グループにアクセスする必要のあるライト要求を受け付けた時、マスタディスクが空いている確率を高めることができるためである。

【0077】以下、実施例の詳細を説明する。

【0078】まず、第1の実施例について詳細に説明す

る。

【0079】図2に示した計算機システムの構成は、第1の実施例にそのまま適用できる。以下、それぞれの詳細について説明する。

【0080】図8は、ディスク装置204の構成である。円板801は、データを記録する媒体であり、1つのディスク装置204には複数存在する回転体である。読み書きヘッド802は、円板801上のデータを読み書きする装置であり、円板308対応に存在する。制御装置インターフェイス803は、制御装置203とのインターフェイスとなる。

【0081】円板801が一回転する間に、読み書きヘッド802がアクセス可能な円状の記録単位をトラック800とよぶ。トラック800は、円板801上に複数存在する。

【0082】図9は、トラック800の構成である。トラック800は、ある位置を基準として、トラック先頭902と、トラック末尾903が定められる。また、トラック800上には、1つ以上のレコード900が存在する。レコード900は処理装置210と制御装置203との間の最小の入出力処理単位である。トラック800上のレコード900の位置は、セル901という、固定長バイトを単位で表現する。(レコード900は、必ず、セル901の先頭から格納開始され、セル901の途中からは、格納開始されない。)セル901の番号は、トラック800の先頭を0番とし、1ずつの昇順につけられる。

【0083】図10は、キャッシュ206の構成である。キャッシュ206は、セグメント1000より構成される。本実施例では、1つのトラック800に対し1つのセグメント1000を割り当て、セグメント1000内には、トラック800全体のデータを格納するものとする。ただし、本発明は、セグメント1000の割当単位を、トラック800に限定する必要はなく、もっと小さい単位、例えば、処理装置210と制御装置203の間のリード/ライト単位であるレコード900としても有効である。

【0084】図11は、ディレクトリ208の構成である。ディレクトリ208は、セグメント管理情報1100、トラック票1101、ならびに、空きセグメント先頭ポインタ1102により構成される。セグメント管理情報1100は、セグメント1000単位に存在する。トラック票1101および空きセグメントポインタ1102は制御装置203内に1つ存在する。

【0085】図12は、セグメント管理情報1100の中にもうける本実施例に必要な情報を示したものである。以下、各パラメータとその内容を示す。

【0086】空きセグメントポインタ1200…トラック800に割り当てていない他のセグメント1000に対応したセグメント管理情報1100へのポインタであ

る。

【0087】キャッシュドトラック番号1201…当該セグメント管理情報500に対応したセグメント400内に格納したディスク装置グループ211の番号、トラック800の番号を表す。

【0088】レコードビットマップ1202…当該セグメント管理情報500に対応したセグメント400内に格納したトラック800上のレコード900の開始位置を表わす。ここで、それぞれのビットはセル901の番号対応に存在するものとする。例えば、レコードビットマップ1202の中のn番目のビットがオンであれば、当該セグメント管理情報1100に対応したn番目のセル901から、レコード301の格納が開始されていることになる。n番目のビットがオフであれば、n番目のセル901から、格納開始されているレコード301は存在しないことになる。図13は、本実施例におけるセグメント1000内の、トラック800上の格納形式を表わしたものである。セグメント1000内には、トラック800上のトラック先頭901から、レコード900が順番に格納される。したがって、レコード900のトラック上で格納開始されるセル901の番号がわかれば、そのレコード301のセグメント1000内の格納開始位置もわかる。

【0089】更新レコードビットマップ1203…当該セグメント管理情報500に対応したセグメント400内に格納されていて、かつ、ライトアフタ処理112が必要なレコード900のビットマップである。ライト処理112が必要なレコード900を以下、ライトアフタレコードと呼ぶ。それぞれのビットは、レコードビットマップ1202と同様、セル901の番号対応に存在する。具体的には、更新レコードビットマップ1203の中のn番目のビットがオンであれば、当該セグメント管理情報1100に対応したn番目のセル901から格納開始されているレコード301が、ライトアフタレコードということになる。更新レコードビットマップ1203は、1台のディスク装置204対応に存在する。具体的に、それぞれの更新ビットマップ1203がどのディスク装置204に対応するかについては、制御用メモリ207の構成の説明の部分で述べる。更新レコードビットマップ1203の領域は、1つのディスク装置グループ211内の定義可能なディスク装置204の数の分だけ用意されている。ただし、使用される更新レコードビットマップ1203の数は、当該ディスク装置グループを構成するディスク装置204の台数である。

【0090】格納済みフラグ1204…当該セグメント管理情報に対応したセグメント1000内に、割当てたトラック800上のレコード900を格納してあるかを示す。

【0091】使用中フラグ1205…当該セグメント管理情報1100に割当てたトラック800に対応した入

出力処理を実行中であることを示す。

【0092】セグメントポインタ1206…当該セグメント管理情報に対応したセグメント1000へのポインタである。

【0093】図14は、トラック票1101、空きセグメント先頭ポインタ1102の構成である。

【0094】トラック票1101は、すべてのディスク装置グループ211のトラック800に関して、そのトラック800に対してセグメント1000が割り当てられているか、いないかを表す。割り当てられている場合には、そのトラック800に割り当てられているセグメント1000に対応したセグメント管理情報1200へのポインタを表す。トラック票1201においては、同じディスク装置グループ211上のトラック800に関する情報は、まとめて、トラック800の番号順に格納される。

【0095】トラック800を割り当ててないセグメント1000に対応したセグメント管理情報1100は、空きセグメント先頭ポインタ1102から順に、空きセグメントポインタ1200で、結合される。結合されているセグメント管理情報1100の集合を空きセグメントキュー1400と呼ぶ。

【0096】図15は、制御情報用メモリ207の構成である。制御情報用メモリ207内には、各ディスク装置グループ211に対応した制御情報であるディスク装置グループ情報1500が含まれる。ディスク装置グループ情報1500の個数は、1台の制御装置203内に定義可能なディスク装置グループ211の数だけ用意されている。

【0097】図16は、ディスク装置グループ情報1500の構成である。

【0098】ディスク装置数1600…当該ディスク装置グループ211内に現在定義されているディスク装置204の数である。

【0099】ディスク装置情報1601…当該ディスク装置グループ情報1500を構成するそれぞれのディスク装置204対応の情報である。ディスク装置情報1600は、1つのディスク装置グループ211内に定義可能なディスク装置204だけ用意される。ただし、有効な情報は、先頭のディスク装置情報1601からディスク装置数1600に定義されている数のディスク装置情報1601までに格納される。ここで、先頭のディスク装置情報1601が、マスタディスクに対応した情報である。

【0100】また、セグメント管理情報1200内のn番目の更新レコードビットマップ1203は、n番目のディスク装置情報1601に対応したディスク装置204である。

【0101】処理装置I/O待ちビット1602…処理装置210から、当該ディスク装置グループ211への

受け付けた入出力要求が待ち状態になっていることを表わすビットである。本ビットの数は、以下のように表わすことができる

処理装置I/O待ちビット1602の数=制御装置203に接続可能な処理装置210の数(ここでは、I台とする。)×1台の処理装置210が1つのディスク装置グループ210に対して並行して処理可能な入出力処理要求の数(ここでは、J個とする。)

したがって、各処理装置210が制御装置203に入出力要求を発行する際には、処理装置210は、制御装置203に以下の2点を通知する。まず、第1点目の内容は、入出力要求を発行する処理装置210が、1番からI番目までの何番の処理装置210であるかという点である。第2点目は、その入出力要求が、指定したディスク装置グループ211への、1番からJ番までの何番目の入出力要求であるかを通知する。

【0102】図17は、ディスク装置情報1601の構成である。

【0103】ディスク装置番号1700…当該ディスク装置情報1700に対応したディスク装置204の番号である。

【0104】処理装置I/O実行中ビット1701…当該ディスク装置情報1700に対応したディスク装置204が、処理装置210からの入出力要求の実行中であることを1ビットで表わす。

【0105】ライトアフト実行中ビット1702…当該ディスク装置情報1700に対応したディスク装置204が、ライトアフト処理112を実行中であることを1ビットで表わす。

【0106】独立ステージ実行中ビット1703…当該ディスク装置情報1700に対応したディスク装置204が、処理装置210とは独立したステージ処理115を実行中であることを1ビットで示す。

【0107】処理装置I/O実行中ビット1701、ライトアフト実行中ビット1702、独立ステージ実行中ビット1703の内同時にオンになるのは高々1つの情報である。また、処理装置I/O実行中ビット1701、ライトアフト実行中ビット1702、独立ステージ実行中ビット1703のすべてがオフであるディスク装置204が、空いた状態にあるディスク装置204ということになる。

【0108】セグメント管理情報ポインタ1704…当該ディスク装置情報1700に対応したディスク装置204で実行中の入出力処理がアクセスするトラック800に割当てたセグメント管理情報1100へのポインタを表わす。

【0109】制御情報用メモリ207内の情報は、電源障害等で消失してしまうと問題があるため、制御情報用メモリ207は、不揮発化しておくことが望ましい。

【0110】制御装置203が実行すべき入出力処理

は、実際には、制御装置203内にそれぞれのディレクタ205が、並行して実行することになる。

【0111】図18には、それぞれのディレクタ205が、本実施例を実行する際に用いる各手続きを示した。以下、それぞれ機能について述べる。

【0112】入出力要求受け付け部1800…処理装置210から受け付けた入出力要求の処理を行う。

【0113】ライトアフト処理スケジュール部1801…ライトアフト処理をスケジュールする。

【0114】独立ステージ処理スケジュール部1802…処理装置210とは独立したステージ処理をスケジュールする。

【0115】ディスク装置転送部1803…ディスク装置204とのリード/ライト転送を実行する部分である。

【0116】図19は、入出力要求受け付け部1800の処理フロー図である。入出力要求受け付け部1800の実行契機は、処理装置210から、新たな入出力要求を受け付けた時である。以下、その処理フローを説明する。

【0117】ステップ1900では、受け取った入出力要求が、ディスク装置204までアクセスする必要がある入出力要求であるかをチェックする。具体的にどのような入出力要求が、ディスク装置204にアクセスする必要があるのかについては、本発明には直接関係しないため、詳細は省略する。受け付けた入出力要求が、ディスク装置204にアクセスする必要のない場合、ステップ1915へジャンプする。

【0118】ステップ1901以降の処理は、受け付けた入出力要求が、ディスク装置204にアクセスする必要がある入出力要求の場合の処理である。

【0119】ステップ1901では、入出力要求がアクセス対象とするトラック300が、セグメント1000を割当て中かを調べる。割当て中であれば、ステップ1903へジャンプする。

【0120】ステップ1902では、入出力要求がアクセス対象とするトラック300にセグメント管理情報1100を割当て、トラック票1101の該当する領域にリンクする。さらに、割当てたセグメント管理情報1100の格納済フラグ1205をオフにし、使用中フラグをオンにする。この時、割当て対象とするセグメント管理情報1100は、空きセグメント先頭キュー1102から空いた状態に有るセグメント管理情報1100を選択する。空いた状態に有るセグメント管理情報1100がない場合には、公知の方法によって現在割当て中のセグメント管理情報を選択することになる。具体的な選択方法は、本発明には関係しないため説明を省略する。この後、ステップ1905へジャンプする。

【0121】ステップ1903では、入出力要求がアクセス対象とするトラック300に割当て中のセグメント

管理情報1100の使用フラグ1205がオンかどうかをチェックする。オンであれば、他の入出力処理がアクセス対象とするトラック300を現在使用中であることになる。したがって、受け付けた入出力要求は、直ちに実行に入れないため、ステップ1914へジャンプする。

【0122】使用中フラグ1205がオフであれば、ステップ1904で、使用中フラグ1205をオンにする。

【0123】次に、ステップ1905で、入出力要求が、リード要求かライト要求かを調べる。本発明では、ディスク装置グループ211にアクセスする必要があるライト要求はマスタディスクにアクセスさせる。したがって、入出力要求がリード要求の場合には、ステップ1908へ分岐する。

【0124】ライト要求の場合、ステップ1906で、マスタディスクが空いた状態にあるかチェックする。このチェックは、入出力対象となっているディスク装置グループ211内のマスタディスクに対応したディスク装置情報1601（すなわち、ディスク装置グループ情報1500内の、先頭のディスク装置情報1601）内の以下の情報をチェックする。すなわち、処理装置I/O実行中ビット1701、ライトアプタ実行中ビット1702、独立ステージ実行中ビット1703のすべてのビットがオフかを（空いた状態にあるかを）チェックする。マスタディスクが、空いた状態にあれば、この後、ステップ1907で、マスタディスクをアクセス対象として選択する。具体的には、マスタディスクに対応するディスク装置情報1601内の処理装置I/O実行中ビット1701オンにする。以上の処理が終了すると、ステップ1910へジャンプして、第1の実施例と同様の処理に入る。

【0125】マスタディスクが空いた状態になれば、当該入出力要求を待ち状態にするため、ステップ1913へジャンプする。

【0126】ステップ1908では、ディスク装置204にアクセスさせる必要があるリード要求に対してディスク装置204を割り当てる。本実施例では、ディスク装置204にアクセスさせる必要があるリード要求に対しては、空いた状態にある任意のディスク装置を割り当てる。したがって、入出力対象となっているディスク装置グループ211の中に空いた状態のディスク装置204があるかチェックする。具体的な処理内容は以下のとおりである。すなわち、処理装置I/O実行中ビット1701、ライトアプタ実行中ビット1702、独立ステージ実行中ビット1703のすべてのビットがオフのディスク装置情報1601を見つける。見つからなかった場合、空いたディスク装置204がないことになり、入出力処理に入れないため、ステップ1913へジャンプする。

【0127】見つかった場合、ステップ1909で、そのディスク装置情報1601内のディスク装置番号1700に対応するディスク装置204をアクセス対象として選択する。具体的には、見つけたディスク装置情報1601内の処理装置I/O実行中ビット1701をオンにする。

【0128】さらに、ステップ1910で、セグメント管理情報ポインタ1704に、当該入出力要求でアクセス対象となるトラック800に割当てたセグメント管理情報1100へのポインタを設定する。

【0129】ステップ1911では、ステップ1909で選択したディスク装置204に対して、位置付け処理要求を発行する。

【0130】ステップ1912では、ディスク装置204の位置付け処理が完了するまで、一度、処理装置210との接続関係を切る処理を、処理装置210との間で実行する。この後、入出力要求受け付け部1800の処理を終了させる。

【0131】空いたディスク装置204がなかった場合、ステップ1913以降で、以下の処理を実行する。

【0132】ステップ1913では、対応するセグメント管理情報1100の使用フラグ1205をオフにする。

【0133】この後、ステップ1914で、格納済フラグ1204がオンかどうかをチェックする。オンの場合ステップ1916へジャンプする。オフの場合、このセグメント管理情報1100に対応したセグメント1000にはデータが入っていないことを示すため、ステップ1915で、このセグメント管理情報を空きセグメントキュー1400に登録する。

【0134】さらに、ステップ1916で、処理装置210に、当該入出力要求に対する処理が、他の入出力処理のために実行に入れなかったということを、ディスク装置グループ情報1500内の処理要求I/O待ちビット1602に設定する。具体的には、処理要求I/O待ちビット1602のどの位置のビットを設定するかを以下の2点から決定し、そのビットの設定を行う。

【0135】まず、第1点は、当該入出力要求を発行した処理装置210が、1番からI番目までの何番の処理装置210であるかということである。

【0136】次に、第2点目は、その入出力要求が指定したディスク装置グループ210への、1番からJ番までの何番の入出力要求であるかということである。

【0137】当該入出力要求がアクセス対象とするトラック800のセグメント管理情報1100が、他の入出力処理によって使用されている場合には、特にセグメント管理情報1100内の情報は操作する必要がない。したがって、ステップ1916へジャンプしてくることになる。

【0138】最後に、ステップ1917で、処理装置2

10に、当該入出力要求に対する処理が、他の入出力処理のために実行に入れないため、待ち状態に入ることと報告する。この後、入出力要求受け付け部1800の処理を終了させる。

【0139】ステップ1918では、ディスク装置204にアクセスする必要のない入出力要求に対して、実行する必要のある処理を実行する。具体的な処理内容は、本発明には直接関係がないため説明を省略する。

【0140】図20は、ライトアфта処理スケジュール部1801の処理フローである。ライトアфта処理スケジュール部1801は、ディレクタ25が空いた時間を利用して実行する。

【0141】ステップ2000では、ライトアфта対象とすべきディスク装置グループ210を決定する。この決定方法は、特に本発明とは、直接関係がないため、説明を省略する。

【0142】ステップ2001では、決定したディスク装置グループの中でマスタディスク以外に、入出力対象とすべき空いた状態にあるディスク装置を見つける。具体的な処理内容は以下のとおりである。すなわち、マスタディスク以外で、処理装置1/O実行中ビット1701、ライトアфта実行中ビット1702、独立ステージ実行中ビット1703のすべてのビットがオフのディスク装置情報1601を見つける。見つからない場合、ライトアфта処理が実行できないため、ライトアфта処理スケジュール部1801の処理を、終了させる。

【0143】見つかった場合、ステップ2002で、ステップ2001で見出したディスク装置情報1601内のライトアфта実行中ビット1702をオンにする。

【0144】ステップ2003では、ステップ2001で見出してディスク装置204にライトアфта処理112が実行可能なトラック800があるかどうかをチェックする。具体的なチェック情報は、トラック票1101から、選択したディスク装置204に対応して更新レコードビットマップ1203中にオンのビットをもつセグメント管理情報1100を探す。さらに、そのセグメント管理情報を他の処理要求を使用中でないことが必要となるため、セグメント管理情報1100内の使用中フラグ1205がオフであるということもライトアфта処理が実行可能な条件となる。見つかった場合、ステップ2005にジャンプする。ライトアфта処理が実行可能なトラック800がない場合、ステップ2004で、ライトアфта実行中ビット1702をオフにする。この後、ライトアфта処理スケジュール部1801の処理を終了させる。

【0145】ステップ2005では、ライトアфта処理すべきトラック800を選択する。ライトアфта処理が実行可能なトラック800が複数存在する場合、どのトラック800を選択するかは本発明には関係しないため、説明を省略する。

【0146】ステップ2006では、ステップ2005で選択したトラック800に対応したセグメント管理情報1100内の使用中フラグ1205をオンにする。

【0147】ステップ2007では、セグメント管理情報ポインタ1704に、当該入出力要求でアクセス対象となるトラック800に割当てたセグメント管理情報1100へのポインタを設定する。

【0148】ステップ2008では、ステップ2001で選択したディスク装置204に対して、位置付け処理要求を発行する。この後、ライトアфта処理スケジュール部1801の処理を終了させる。

【0149】図21は、独立ステージ処理スケジュール部1802の処理フローである。独立ステージ処理スケジュール部1802も、ディレクタ25が空いた時間を利用して実行する。

【0150】ステップ2100では、処理装置210とは独立したステージ処理115を実行すべきディスク装置グループ210を決定する。この決定方法も、ステップ2000と同様、特に本発明とは、直接関係ないため、説明を省略する。

【0151】ステップ2101では、ステップ2100で見出したディスク装置グループ211内に、処理装置210とは独立なステージ処理を実行すべきトラック800があるかどうかをチェックする。このチェック方法も本発明とは直接関係ないため、説明を省略する。実行すべきトラック800がない場合、独立ステージ処理スケジュール部1802の処理を終了させる。

【0152】ステップ2102では、処理装置210とは独立したステージ処理の対象とするトラック800を選択する。処理装置210とは独立したステージ処理が実行可能なトラック800が複数存在する場合、どのトラック800を選択するかは本発明には関係しない。したがって、その説明を省略する。

【0153】ステップ2103では、ステップ2004で選択してトラック800へ、セグメント管理情報1100を割当てる。（処理装置210とは独立したステージ処理115を実行すべきトラック800はキャッシュ206内にステージされていないトラック800である。）この割当て方法は、ステップ1902で示したとおりである。さらに、割当てたセグメント管理情報1100内の、格納済みフラグ1204をオフに、使用中フラグ1205をオンにする。

【0154】ステップ2104では、ステップ2100で決定したディスク装置グループの中で入出力対象とすべき空いた状態にあるディスク装置を見つける。具体的な処理内容はステップ1908と同様であるため、説明を省略する。見つからない場合、処理装置210とは独立なステージ処理115が実行できない。したがって、ステップ2105で、割り当てたセグメント管理情報1100を、空きセグメントキュー1400に戻す。この

後、独立ステージ処理スケジュール部1802の処理を、終了させる。

【0155】見つかった場合、ステップ2106で、ステップ2103で見出したディスク装置情報1601内の独立ステージ実行中ビット1702をオンにする。

【0156】ステップ2107では、セグメント管理情報ポインタ1704に、当該入出力要求でアクセス対象となるトラック800に割当てたセグメント管理情報1100へのポインタを設定する。

【0157】ステップ2108では、ステップ2001で選択したディスク装置204に対して、位置付け処理要求を発行する。この後、独立ステージ処理スケジュール部1802の処理を終了させる。

【0158】図22は、ディスク装置転送部1803の処理フロー図である。ディスク装置転送部1803の実行契機は、ディレクタ205がディスク装置204の位置付け完了報告を受け取った時である。

【0159】ステップ2200では、当該ディスク装置204に対応したディスク装置情報1601内のセグメント管理情報ポインタ1704がポイントするセグメント管理情報1100を処理対象として選択する。以下、単にセグメント管理情報1100と述べた場合には、ステップ2200で選択したセグメント管理情報1100を指す。また、セグメント管理情報1100内の情報を単に示した場合、ステップ2200で選択したセグメント管理情報1100内の情報を指す。

【0160】ステップ2201では、当該ディスク装置に204に対応するディスク装置情報1601内の処理装置I/O実行中ビット1701をオンかどうかをチェックする。オフであれば、実行中の入出力処理が、処理装置210から受け付けた入出力要求に対する処理でないことを示すため、ステップ2212へジャンプする。

【0161】処理装置I/O実行中ビット1701をオンの場合には、実行中の入出力処理が、処理装置210から受け付けた入出力要求に対する処理であることを示す。したがって、ステップ2202で、処理装置210に位置付け処理が完了したことを報告し、再び接続状態にはいる。

【0162】ステップ2203では、処理装置から受け取った入出力要求がリード要求かライト要求であるかを判別する。リード要求の場合、ステップ2209へジャンプする。

【0163】ライト要求の場合、ステップ2204で、処理装置204から受け取ったデータを、ディスク装置204と処理対象として選択したセグメント管理情報1100に対応したセグメント1000に書き込む。この時、実際にデータを書き込んだディスク装置204上のレコード800のセル901の番号を認識し以下の処理を行う必要がある。まず、セグメント1000内に書き込むデータもその認識したセル901に対応した位置に

書き込む。さらに、処理対象として選択したセグメント管理情報1100内のマスタディスク以外のすべてのディスク装置204に対応した更新レコードビットマップ1203に対して以下の処理を行う。すなわち、更新レコードビットマップ1203の中で、上記で認識したセル901の番号に対応したビットをオンにする。さらに、この後処理装置210に入出力処理が完了したことを報告する。

【0164】ステップ2205では、処理対象としているセグメント管理情報1100内の格納済フラグ1204がオンかをチェックする。オンの場合、処理対象としているトラック300上のレコード900は、セグメント1000内にステージされているため、ステップ2215へジャンプする。

【0165】格納済フラグ1204がオフの場合、処理対象としているトラック300上のレコード900は、セグメント1000内にステージされていないことになる。したがって、ステップ2206以降で以下の処理を行う。ステップ2206では、ステップ2204で認識したセル901の番号に対応するレコードビットマップ1202のビット位置をオンにする。

【0166】次に、ステップ2207では、処理中のトラック800の残りのレコード900をセグメント1000内にステージする処理を実行する。この場合も、ステージ対象とするレコード900のセル901の番号を認識し以下の処理を行う必要がある。まず、セグメント1000内にステージするレコード900もその認識したセル901に対応した位置に書き込む。さらに、認識したセル901の番号に対応するレコードビットマップ1202のビット位置をオンにする。この後、ステップ2208で、格納済フラグ1204をオンにして、ステップ2215へジャンプする。

【0167】ステップ2209以下では、処理装置210から受け付けたリード要求の処理を行う。

【0168】ステップ2209では、処理対象とするセグメント管理情報1100内の格納済フラグ1204がオンかをチェックする。格納済フラグ1204がオンの場合、すでにレコード900が、セグメント1000内に格納されている。したがって、ステップ2210で要求されたレコード900をディスク装置204上から処理装置210に送る。この後、処理装置210に入出力処理が完了したことを報告する。次に、ステップ2215へジャンプする。

【0169】格納済フラグ1204が、オフの場合、処理中の処理対象としているトラック300上のレコード900は、セグメント1000内にステージされていない。したがって、ステップ2211以降で以下の処理を行う。まず、ステップ2210では、要求されたレコード900をディスク装置204上から処理装置210に送ると共に、セグメント1000にもステージする。こ

の場合も、ステージ対象としたレコード900のセル901の番号を認識し以下の処理を行う必要がある。まず、セグメント1000内にステージするレコード900もその認識したセル901に対応した位置に書き込む。さらに、処理対象として選択したセグメント管理情報1100内のレコードビットマップ1202の中の、認識したセル901の番号に対応するビットをオンにする。この後、処理装置210に入出力要求を完了したことを報告する。次に、処理対象としているトラック300上の残りのレコード900をステージするために、ステップ2207へジャンプする。

【0170】ステップ2212では、当該ディスク装置に204に対応するディスク装置情報1601内のライトアフタ実行中ビット1702がオンかどうかをチェックする。オフの場合、処理装置210とは独立なステージ処理115を実行するためにステップ2214へジャンプする。

【0171】ステップ2213では、処理対象とするセグメント管理情報1100中の更新レコードビットマップ1203によって、すべてのライトアフタレコードを認識する。さらに、認識したすべてのライトアフタレコードをディスク装置204上に書き込む。この後、当該ディスク装置204に対応する更新レコードビットマップ1203をすべて0クリアする。次に、ステップ2215へジャンプする。

【0172】ステップ2214では、処理装置210とは独立なステージ処理を実行する。具体的には、処理対象としているトラック800上のすべてのレコード900をセグメント1000内にステージする。この場合も、ステージ対象としたレコード900のセル901の番号を認識し以下の処理を行う必要がある。まず、セグメント1000内にステージするレコード900もその認識したセル901に対応した位置に書き込む。さらに、処理対象とするセグメント管理情報1100内のレコードビットマップ1202に対して以下の処理を行う。すなわち、認識したセル901の番号に対応するレコードビットマップ1202のビット位置をオンにする。加えて、格納済フラグ1204をオンにする。

【0173】ステップ2215以下では、終端処理として以下の処理を行う。

【0174】まず、ステップ2215では、処理対象としてはセグメント管理情報1100内の使用中フラグ1205をオフにする。次に、ステップ2216では、当該ディスク装置204に対応するディスク装置情報1601内の、処理装置I/O実行中ビット1701、ライトアフタ実行中ビット1702、独立ステージ実行中ビット1703のすべてのビットをオフする。

【0175】最後に、ステップ2217では、処理対象としているディスク装置グループ211に対応する、処理要求I/O待ちビット1602がオンになっている入

出力要求の待ち状態を解放するために以下の処理を行う。すなわち、オンになっているビットから、1番からI番までの処理装置210、および1番からJ番までの入出力番号によって決定されるすべての入出力要求の待ち状態を解放する。具体的には、再び、それらの入出力要求を発行するようそれぞれの処理装置に通知する。この後、ディスク装置転送部1802の処理を終了させる。

【0176】次に、第2の実施例について説明する。第2の実施例が第1の実施例と異なる点は、以下のとおりである。

【0177】ディスク装置グループ211にアクセスする必要のあるリード要求、処理装置210とは独立なステージ処理は、マスタディスク以外のディスク装置204を優先して選択する点である。

【0178】第1の実施例において、図8から図17までに示したそれぞれのデータ構成は、第2の実施例でもそのまま用いることができる。

【0179】図18に示した、ディレクタ205内で第1の実施例を実行するために必要なモジュールの構成も、第3の実施例にそのまま適用できる。しかし、それぞれのモジュールの処理フローに関しては、入出力要求受け付け部1800、独立ステージ処理スケジュール部1802が、第1の実施例と若干ことなる。ただし、他のモジュールの処理フローに関しては、第1の実施例の処理フローがそのまま適用できる。

【0180】図23は、第3の実施例における入出力要求受け付け部1800の処理フローである。入出力要求受け付け部1800の実行契機は、第1の実施例の場合と同様である。

【0181】以下、図19に示した第1の実施例における処理フローと図23の処理フローの相違点について述べる。なお、図23の処理フローにおいて、図19の処理フローと処理内容がまったく同じ部分に関しては、ステップ番号を等しくしてある。図23と、図19の処理フローとの差異は、図19のステップ1908の代わりに、ステップ2300が入っている点である。

【0182】ステップ2300では、マスタディスク以外のディスク装置204が空いているかを優先的に選択している。これは、第2の実施例が、ディスク装置グループ211にアクセスする必要のあるリード要求、マスタディスク以外のディスク装置204を優先して選択するためである。具体的には、マスタディスク以外のディスク装置204のディスク装置情報1601内の以下の情報をチェックする。すなわち、処理装置I/O実行中ビット1701、ライトアフタ実行中ビット1702、独立ステージ実行中ビット1703のすべてのビットがオフかをチェックする。

【0183】空いたディスク装置204があれば、そのディスク装置204をアクセス対象として選択するため

に、ステップ1909へジャンプして、図19の処理フローと同様の処理に入る。空いたディスク装置204がない場合、マスタディスクの空きをチェックするため、ステップ1906へジャンプして、図19の処理フローと同様の処理に入る。

【0184】以下の点以外は、図23の処理フローと第19の処理フローはまったく同様であるため、説明を省略する。

【0185】図7は、第2の実施例における独立ステージ処理スケジュール部1802の処理フローである。独立ステージ処理スケジュール部1802の実行契機は、第1の実施例の場合と同様である。

【0186】以下、図21に示した第1の実施例における処理フローと図7の処理フローの相違点について説明する。なお、図7の処理フローにおいて、図21の処理フローと処理内容がまったく同じである部分に関しては、ステップ番号を等しくしてある。

【0187】図7の処理フローが、図21に示した第1の実施例における処理フローと異なる点は以下の点である。

【0188】まず、ステップ2102の後、空いたディスク装置204を見つける際に、第2の実施例では、ステップ2400において、マスタディスク以外の空いたディスク装置204を見つけている点が異なる。これは、第2の実施例が、処理装置210とは、独立なステージ処理は、マスタディスク120以外のディスク装置204を優先して選択するためである。具体的な処理内容は、ステップ2300と同様であるため、説明を省略する。

【0189】空いたディスク装置204があれば、そのディスク装置204をアクセス対象ディスクとして選択するために、ステップ2104へジャンプして、第1の実施例と同様の処理に入る。

【0190】マスタディスク以外のディスク装置204が空いた状態になれば、ステップ2401で、マスタディスクが空いた状態にあるかをチェックする。以上の処理は、ステップ1906の処理と同様であるため説明を省略する。マスタディスクが空いていれば、ステップ2402で、マスタディスクをアクセス対象として選択する。具体的には、マスタディスクに対応するディスク装置情報1601内の処理装置I/O実行中ビット1701オンにする。この後、ステップ2107へジャンプして、第1の実施例と同様の処理に入る。

【0191】マスタディスクが空いていない場合、処理装置210とは、独立なステージ処理は実行できない。したがって、ステップ2105へジャンプして、第1の実施例と同様の処理に入る。以上の点以外は、図7の処理フローと第21の処理フローはまったく同様であるため、説明を省略する。

【0192】

【発明の効果】本発明によれば、1台以上のディスク装置からなるディスク装置グループに同一のデータを書き込む機能をもつキャッシュを有する制御装置の高性能化／高信頼化がバランスよく実現できる。これは、本発明により、信頼性を損なわない範囲で、ディスク装置の間で入出力処理の分散が可能となり、制御装置の実行可能な入出力処理の並列度が向上できるためである。

【図面の簡単な説明】

【図1】本発明の第1の実施例における制御装置203の基本動作である。

【図2】本発明の対象となる計算機システムの構成である。

【図3】処理装置210から受け付けた入出力要求と、処理装置210からの入出力要求とは、制御装置203が独立に実行する入出力処理の並列動作を表している。

【図4】複数の処理装置210から、受け付けた複数の入出力要求どうしの並列動作を表している。

【図5】1つの処理装置210から、受け付けた複数の入出力要求どうしの並列動作を表している。

【図6】本発明の第2の実施例における制御装置203の基本動作である。

【図7】第2の実施例における独立ステージスケジュール処理部1802の処理フロー図を表す。

【図8】ディスク装置24の構成である。

【図9】トラック800の構成である。

【図10】キャッシュ206の構成である。

【図11】ディレクトリ208の中にもうける本発明で必要な情報である。

【図12】セグメント管理情報1100の中にもうける本発明で必要な情報である。

【図13】トラック票1101、空きセグメントキュー先頭ポインタ1102の構成である。

【図14】セグメント1000内のトラック800上のレコード901の格納形式を表す。

【図15】制御用メモリ207上に格納する情報を表す。

【図16】ディスク装置グループ情報1500の構成を表す。

【図17】ディスク装置情報1601の構成を表す。

【図18】本発明に関係するディレクトリ205内のモジュールを表す。

【図19】入出力要求受け付け部1800の処理フロー図を表す。

【図20】ライトアフトスケジュール処理部1801の処理フロー図を表す。

【図21】独立ステージスケジュール処理部1802の処理フロー図を表す。

【図22】ディスク装置リードライト処理部1803の処理フロー図を表す。

【図23】第2の実施例における入出力要求受け付け部



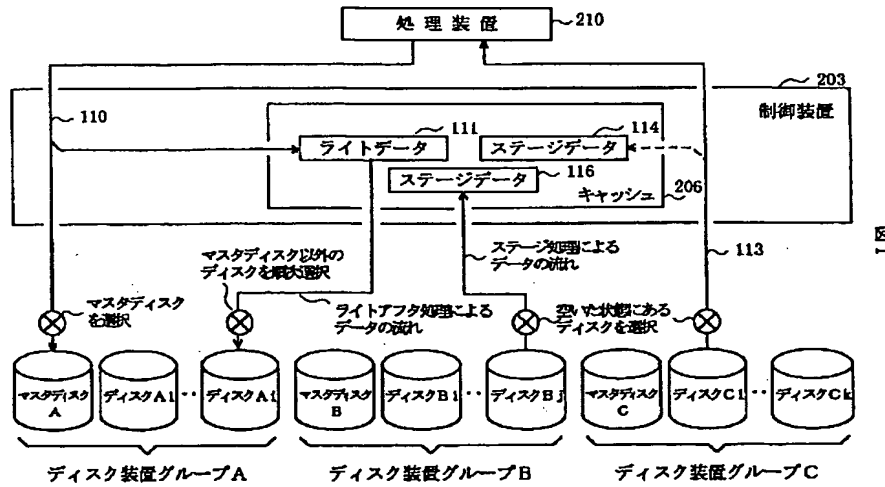
1800の処理フロー図を表す。

【符号の説明】

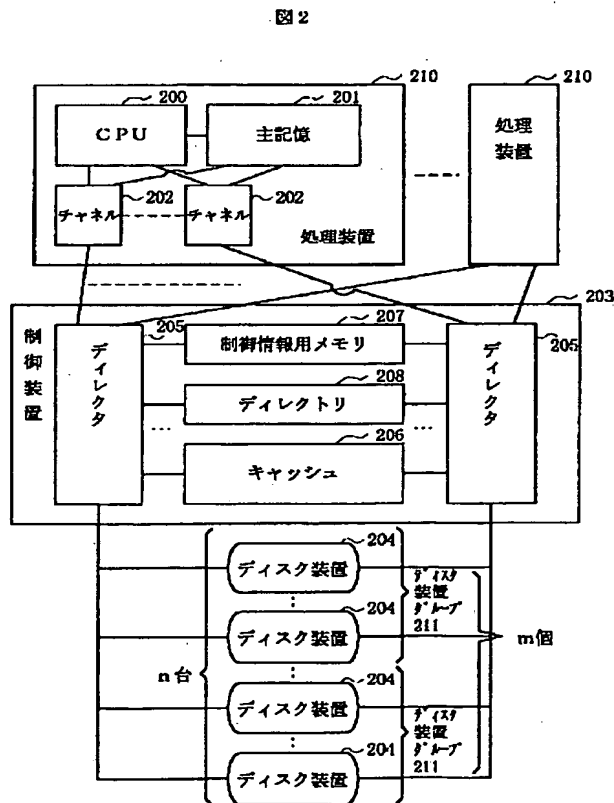
203…制御装置、210…処理装置、211…ディス

ク装置グループ、110…ライト要求、113…リード要求、112…ライトアプタ処理、115…ステージ処理、120…マスタディスク。

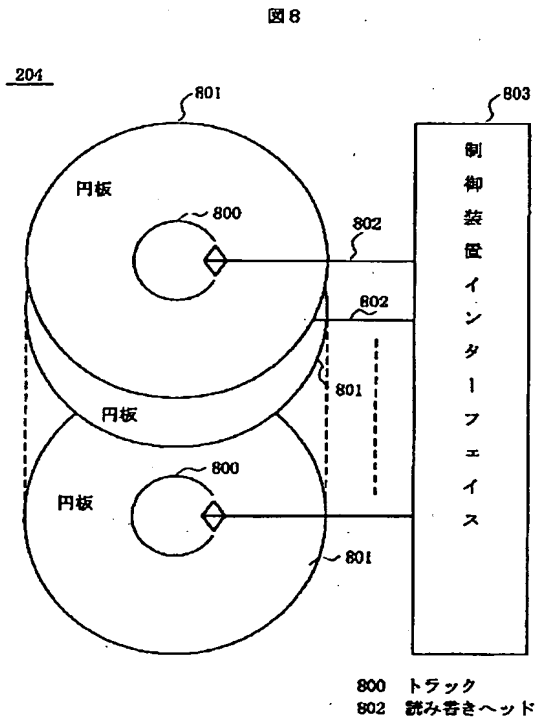
【図1】



【図2】



【図8】



【図 3】

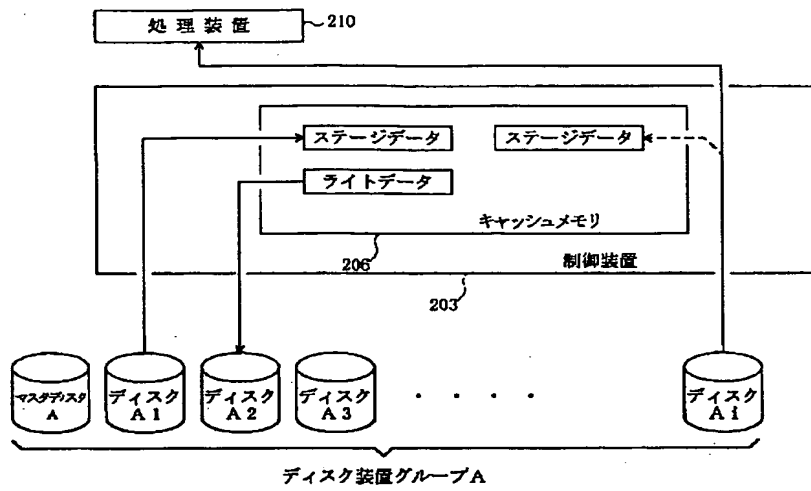


図 3

【図 4】

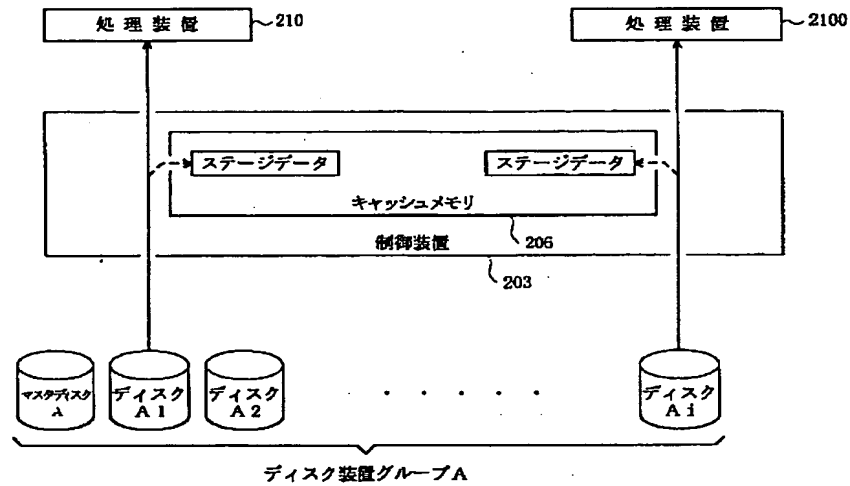


図 4

【図 5】

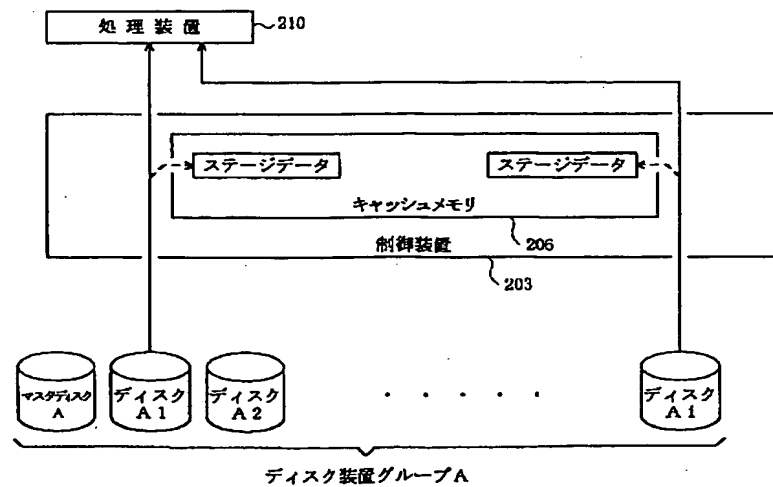
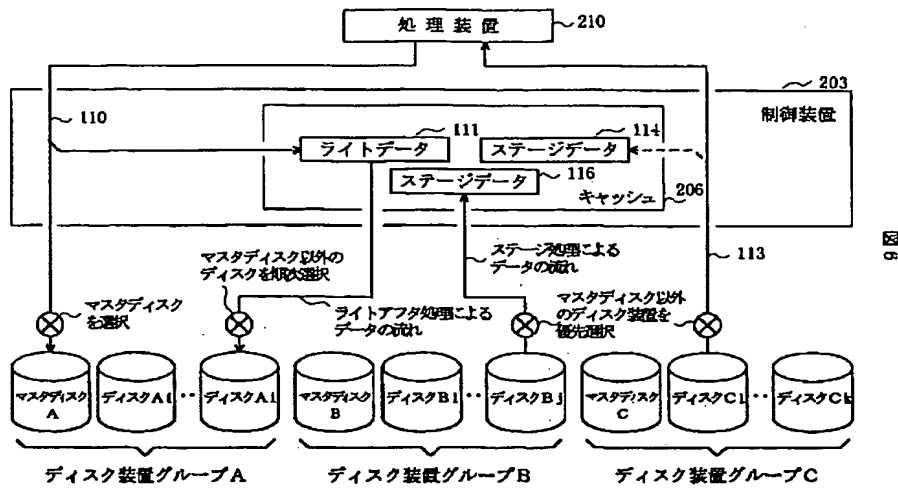
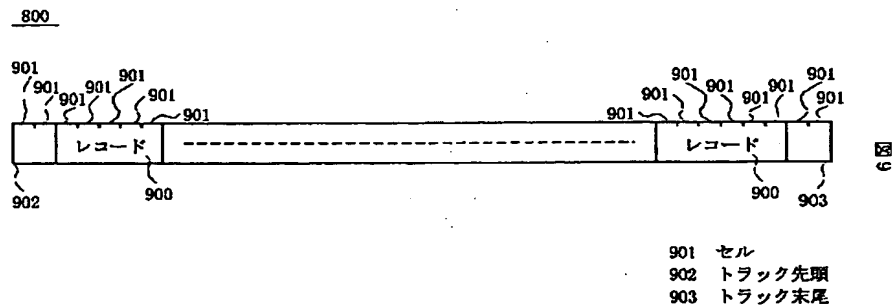


図 5

【図6】

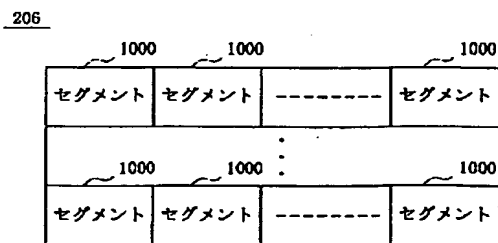


【図9】



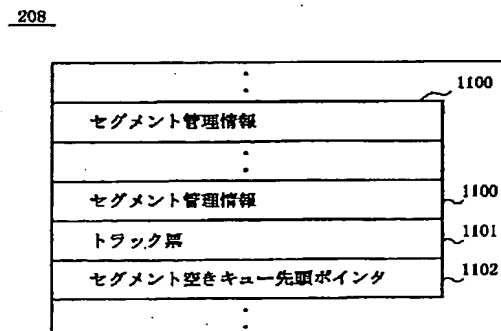
【図10】

図10



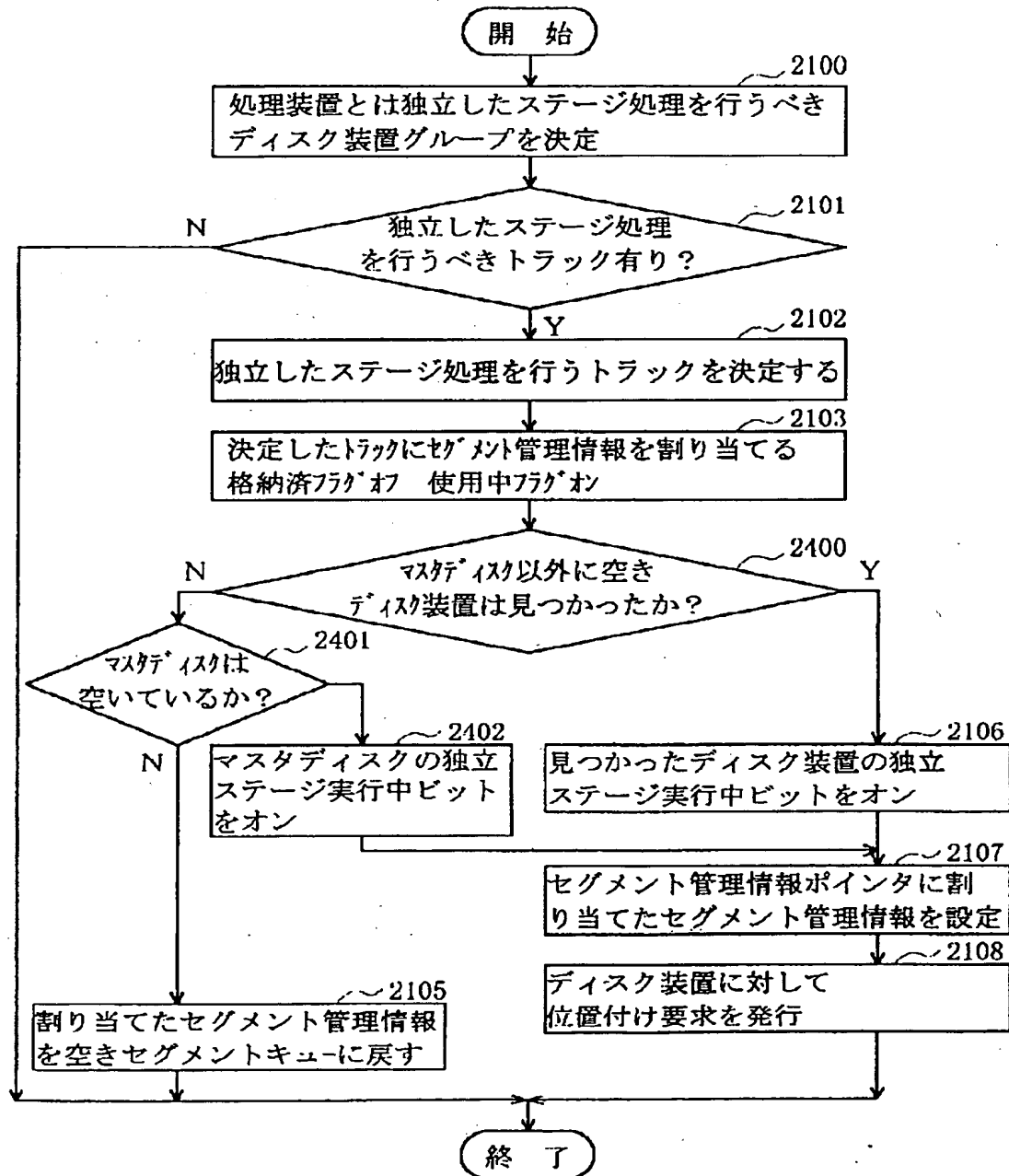
【図11】

図11



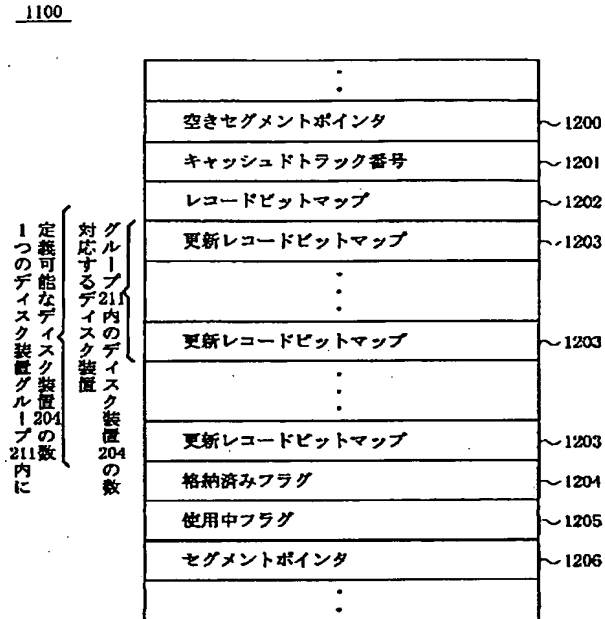
【図7】

図7



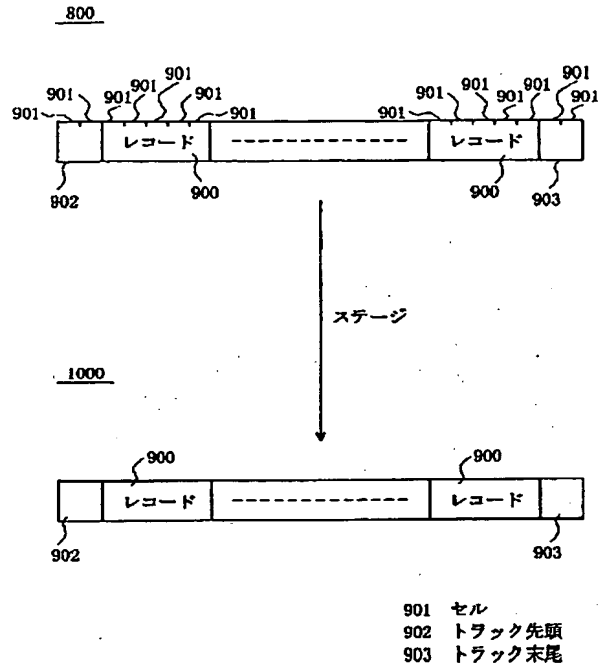
【図12】

図12



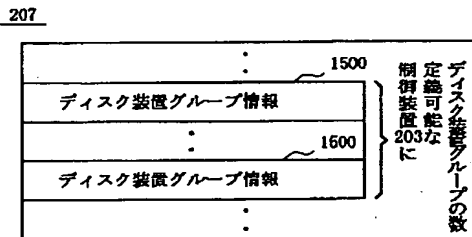
【図13】

図13



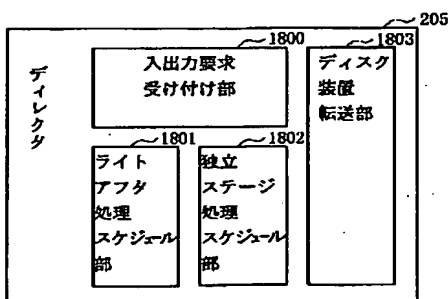
【図15】

図15



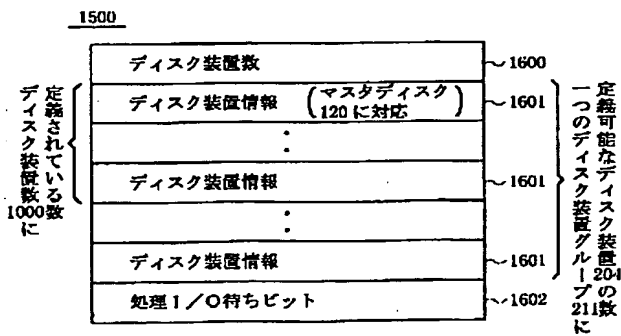
【図18】

図18



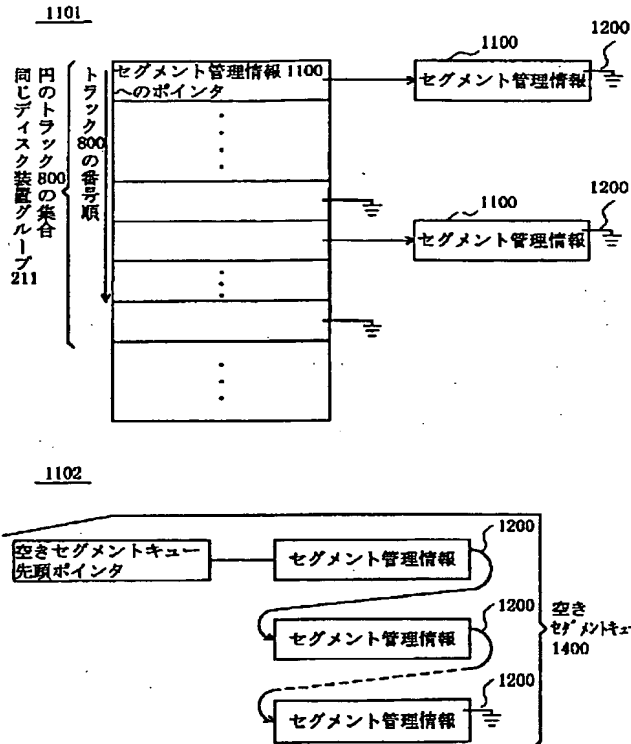
【図16】

図16



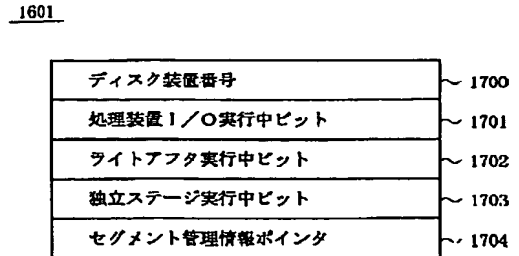
【図 14】

図 14



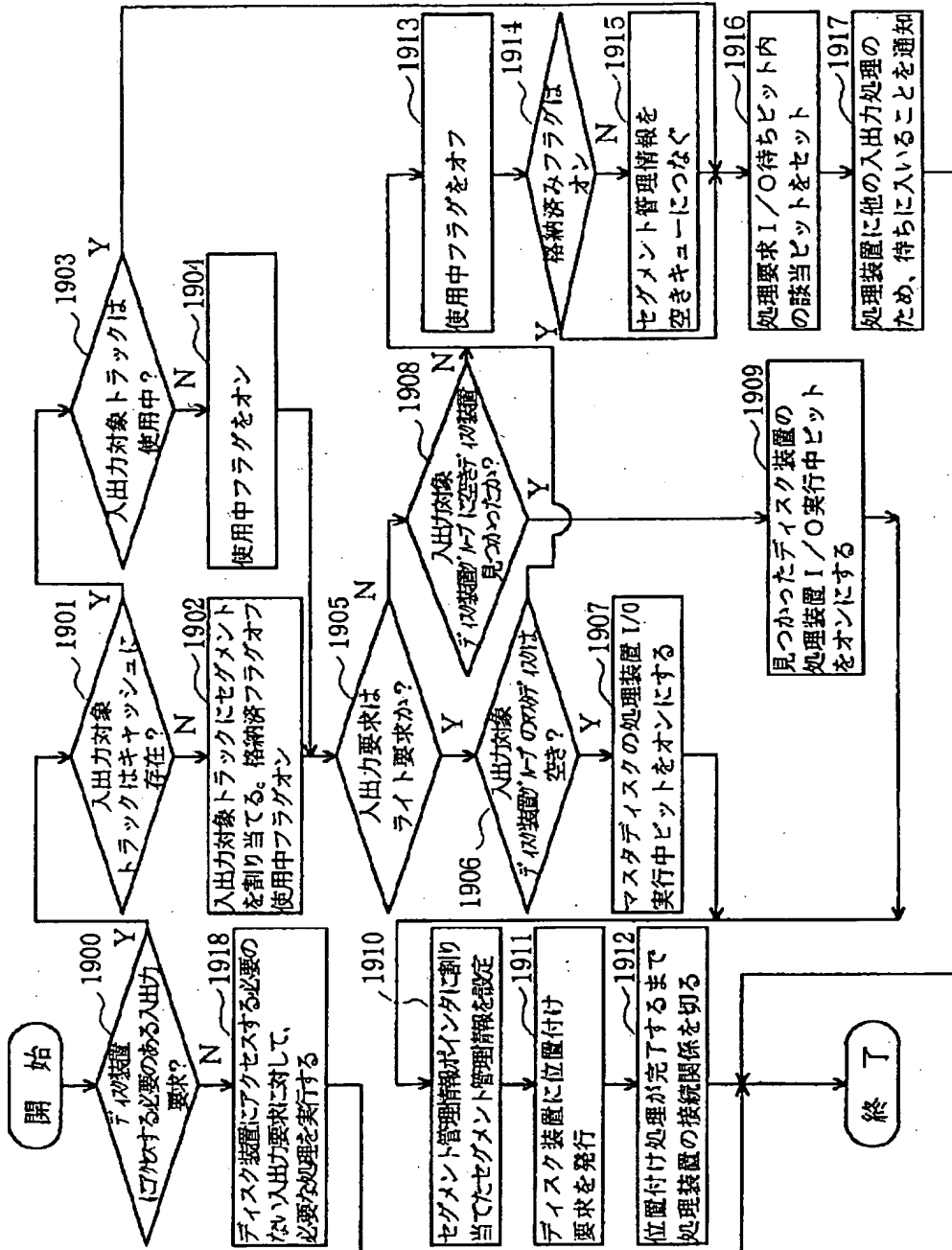
【図 17】

図 17



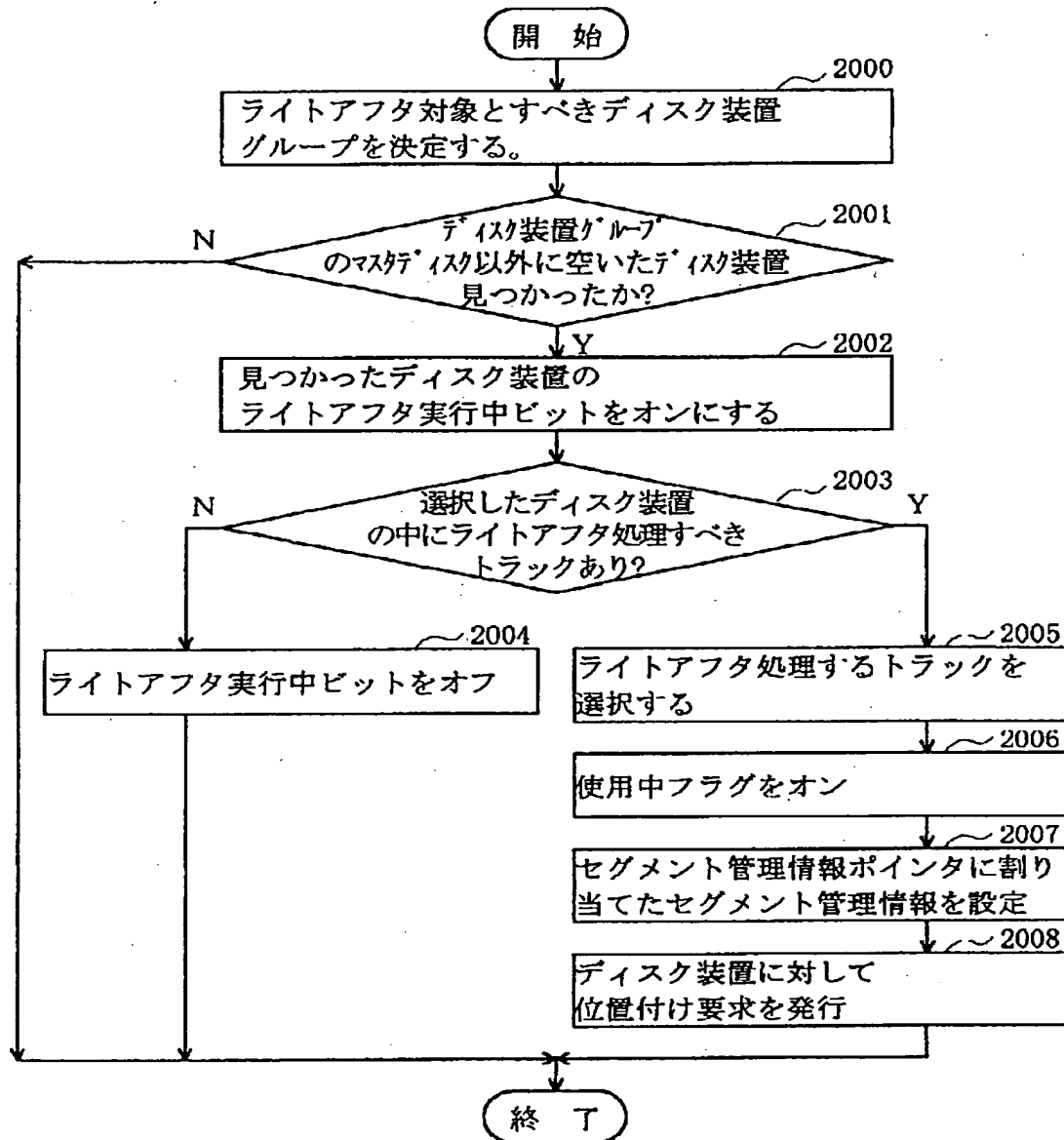
【図 19】

図 19



【図20】

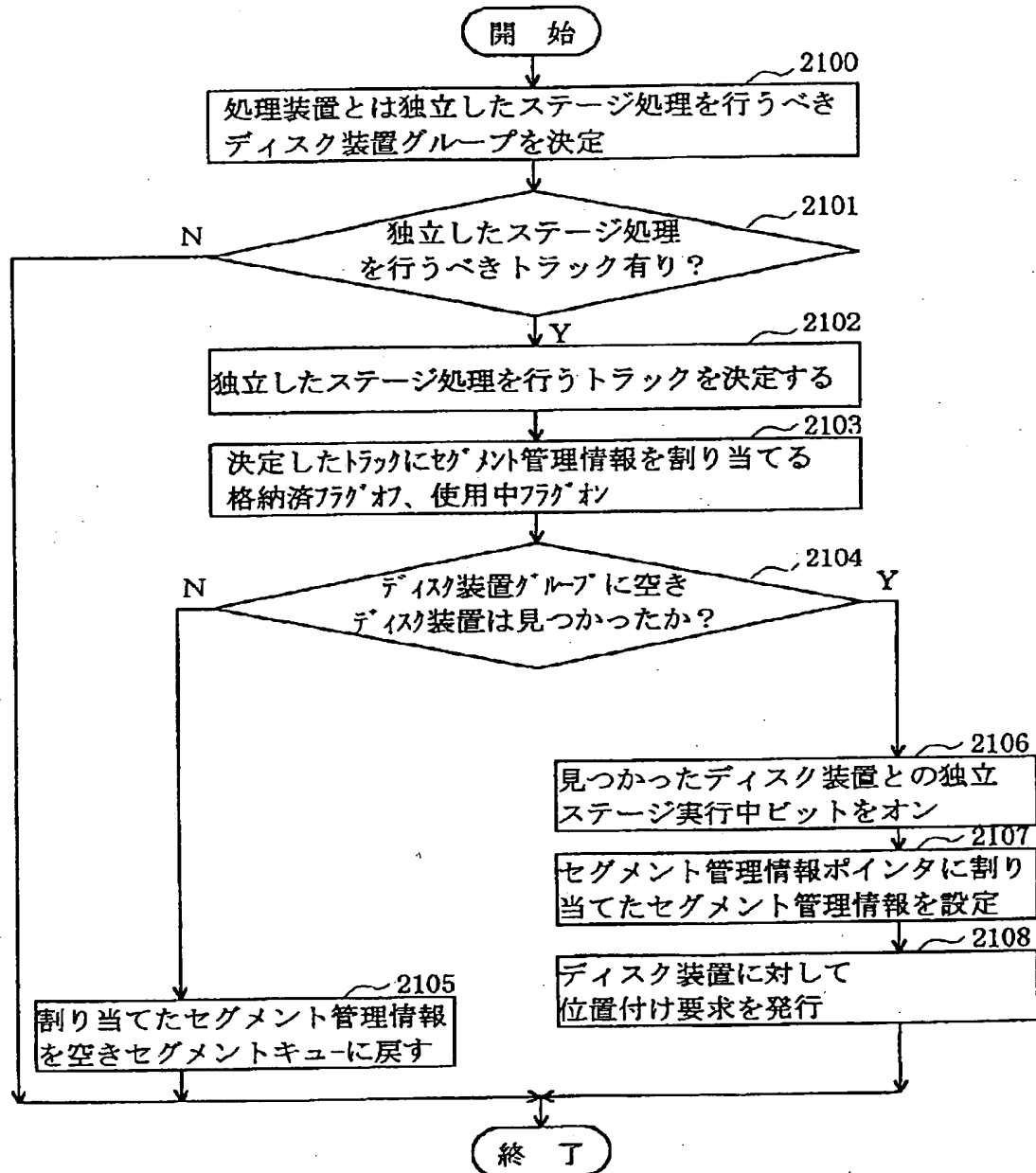
図20





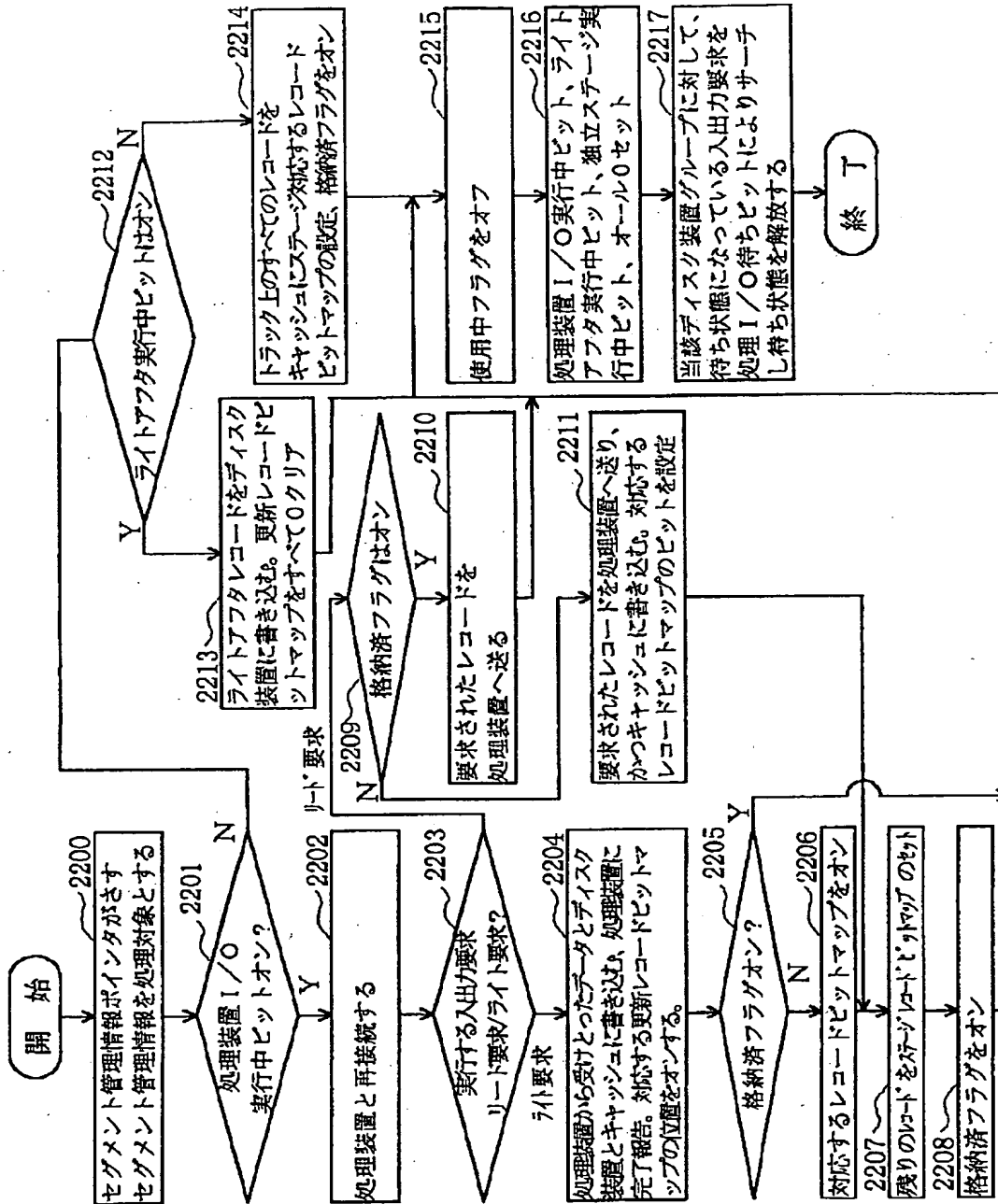
【図21】

図21



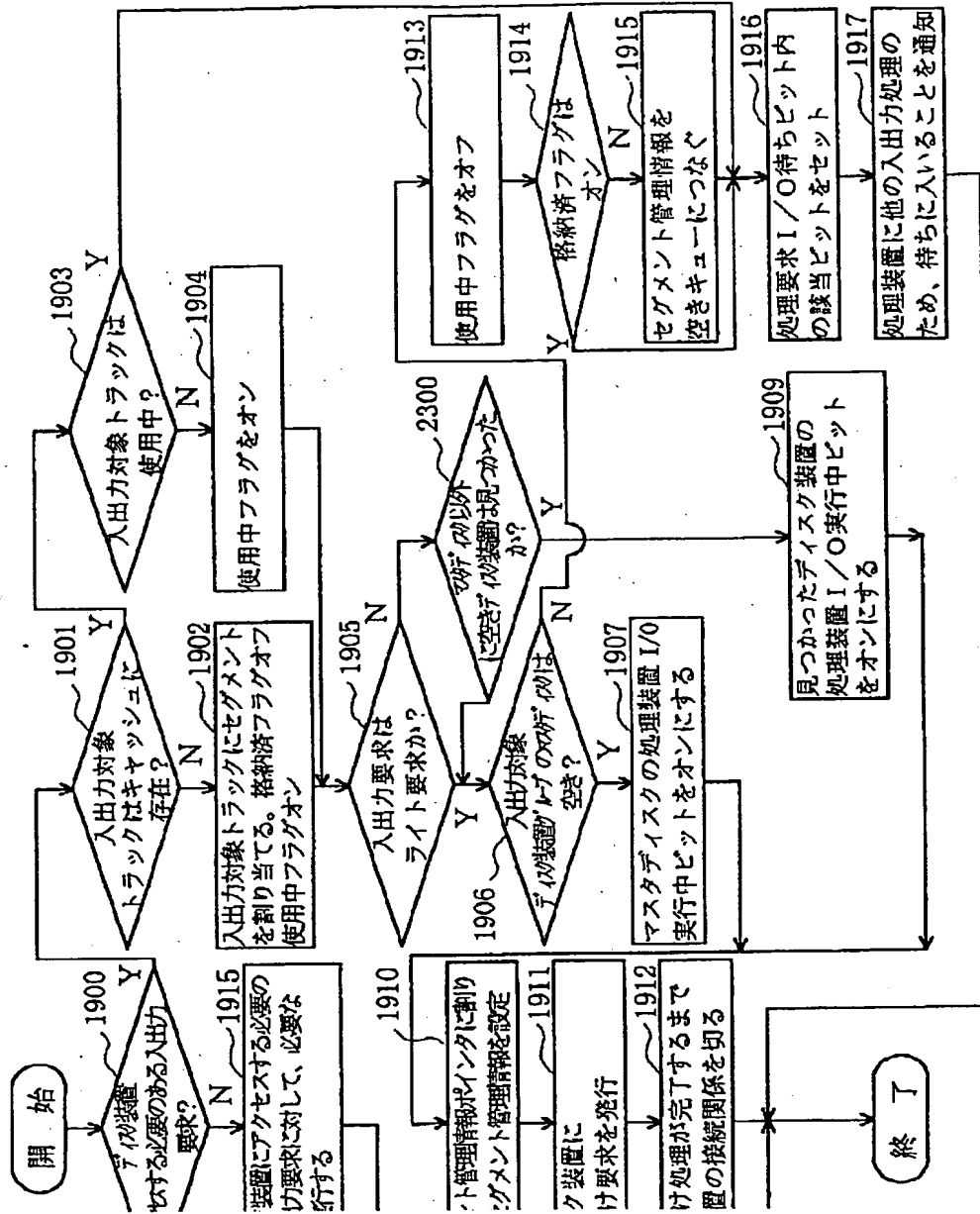
【図22】

図22



【図23】

図23



フロントページの続き

(72)発明者 安積 義弘  
 神奈川県小田原市国府津2880番地 株式会  
 社日立製作所小田原工場内

(72)発明者 桑原 善祥  
 神奈川県小田原市国府津2880番地 株式会  
 社日立製作所小田原工場内

(72)発明者 北嶋 弘行  
神奈川県川崎市麻生区王禅寺1099番地 株  
式会社日立製作所システム開発研究所内